Crop Yield Prdiction

Abstract:

This project aims to develop a robust predictive model for accurately estimating crop yield across various types of crops. Utilizing machine learning techniques, the model will analyze a range of variables, including climatic conditions, soil properties, agricultural practices, and historical yield data. The primary goal is to provide farmers, agronomists, and policymakers with actionable insights to optimize crop production and manage resources efficiently. By leveraging advanced data analytics, this project seeks to enhance agricultural productivity and sustainability, addressing the critical need for food security in the face of growing global population and climate change challenges

Introduction:

Agriculture remains the backbone of many economies worldwide, with crop yield being a critical factor in ensuring food security and economic stability. Accurate prediction of crop yield is essential for efficient resource management, strategic planning, and policy formulation. Traditional methods of estimating crop yield often rely on historical data and expert intuition, which can be time-consuming and prone to inaccuracies due to the complex interplay of multiple influencing factors.

In recent years, the advent of machine learning and big data analytics has opened new avenues for improving the precision of crop yield predictions. These technologies enable the analysis of vast and diverse datasets, encompassing weather patterns, soil characteristics, farming practices, and crop health indicators. By harnessing these data-driven techniques, it is possible to create predictive models that can provide timely and accurate forecasts, helping farmers to make informed decisions about planting, irrigation, fertilization, and harvesting.

The primary objective of this project is to develop a predictive model that integrates various data sources to estimate crop yields accurately for multiple crop types. The model will employ machine learning algorithms to identify patterns and correlations within the data, enabling it to predict future yields based on current and historical inputs. This predictive capability is expected to enhance agricultural productivity by allowing for better planning and optimization of farming practices.

To achieve this, the project will follow a structured approach, starting with data collection and preprocessing, followed by the selection and training of appropriate machine learning models. The performance of these models will be evaluated using relevant metrics, and the best-performing model will be deployed for practical use. Ultimately, this project aims to contribute to the broader goal of sustainable agriculture by providing tools that support efficient crop management and resource allocation.

1. Problem Statement:

To develop a predictive model that accurately estimates the crop yield for various crops cultivated

2. Feature description:

This dataset encompasses agricultural data for multiple crops cultivated across various states in India from the year **1997 till 2020**. The dataset provides crucial features related to crop yield prediction, including crop types, crop years, cropping seasons, states, areas under cultivation, production quantities, annual rainfall, fertilizer usage, pesticide usage, and calculated yields.

Columns Description:

1. **Crop**: The name of the crop cultivated.
2. **Crop_Year**: The year in which the crop was grown.
3. **Season**: The specific cropping season (e.g., Kharif, Rabi, Whole Year).
4. **State**: The Indian state where the crop was cultivated.

5. **Area**: The total land area (in hectares) under cultivation for the specific crop.
6. **Production**: The quantity of crop production (in metric tons).
7. **Annual_Rainfall**: The annual rainfall received in the crop-growing region (in mm).
8. **Fertilizer**: The total amount of fertilizer used for the crop (in kilograms).
9. **Pesticide**: The total amount of pesticide used for the crop (in kilograms).
10. **Yield**: The calculated crop yield (production per unit area).

3. Exploratory Data Analysis:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

3.1 Year wise analysis:

- In this, for Yield Variation Over Years,The graph show how the total yield of crops has varied over the years covered by the dataset. The total yield of crops increases over the years.
- For Area under cultivation Over Years, Similar to the previous plot, each data point on the graph represents the total area under cultivation for a specific year.
- For Use of fertilizer Over Years, Similar to the previous plots, each data point on the graph represents the total fertilizer usage for a specific year. Use of fertilizer increase over the year

3.2 Crop wise analysis:

- Yield Variation for Top 20 Crops: Highest yield producing crop is rice second highest is maize
- Fertilizer used for Top 20 Crops: Fertilizer used for rice crop is more than the fertilizer used for wheat and the rest of all crops
- Area under cultivation: Similarly as above rice crop has more cultivation area then wheat and then rest of the crops

3.3 Season Wise analysis:

- Total Percentage of Each Season Category:
  Kharif season is 51.59%
  rabi season is 36.30%
  Summer season is 6.99%
  autumn season is 2.83%
  winter season is 2.28%

3.4 State wise Annual Rainfall Analysis:

According to the graph the highest annual rainfall is for Chhattisgarh, second highest is West Bengal and Third hightest is Karnataka

4. Feature Engineering:

- Label Encoding:
  LabelEncoder converts the categorical text data in 'Crop' and 'State' columns into numerical values.
- Standardization:
  standardScaler removes the mean and scales the features to unit variance (mean = 0, standard deviation = 1).

5.Cross validation:

Techniques Used:

- K-Fold Cross-Validation:
  K-Fold Cross-Validation is a resampling procedure used to evaluate the performance of a model. It divides the dataset into k equally sized "folds".The model is trained on k−1 folds and tested on the remaining fold. This process is repeated k times, with each fold being used as the test set exactly once.
  The results from each fold are then averaged to provide an overall performance metric.
- K-Nearest Neighbors (KNN) Classifier:
  KNN is a simple, instance-based learning algorithm where predictions for new data points are made by finding the k most similar instances in the training data and taking the majority vote (for classification) or average (for regression) of their outcomes.In this code, the number of neighbors (k) is set to 5 (n_neighbors=5).
- Output:

```
TRAIN: [ 2164  2165  2166 ... 10814 10815 10816] TEST: [   0
1    2 ... 2161 2162 2163]
Fold accuracy: 81.33%
TRAIN: [    0     1     2 ... 10814 10815 10816] TEST: [2164
2165 2166 ... 4325 4326 4327]
```

```
Fold accuracy: 81.42%
TRAIN: [    0     1     2 ... 10814 10815 10816] TEST: [4328
4329 4330 ... 6488 6489 6490]
Fold accuracy: 82.20%
TRAIN: [    0     1     2 ... 10814 10815 10816] TEST: [6491
6492 6493 ... 8651 8652 8653]
Fold accuracy: 81.55%
TRAIN: [   0    1    2 ... 8651 8652 8653] TEST: [ 8654  8655
8656 ... 10814 10815 10816]
Fold accuracy: 80.81%
Cross-Validation Average Accuracy: 81.46
```

This output indicates the performance of the KNN model across the 5 folds and provides an overall average accuracy, giving a robust estimate of the model's performance on unseen data.

6. Model Building:

- Split the data into training and testing sets. We split train and test set in 70 and 30 proposition .
- Applying different algorithms
    1. Decision Tree Classifier
    2. Gaussian Naïve Bayes Classifier
    3. Random Forest.
    4. Logistic Regression

Output :

1. Decision Tree Classifier
```
Accuracy of Decision Tree: 67.46
```
2. Gaussian Naïve Bayes Classifier
```
Accuracy of Gaussian Naive Bayes: 60.02
```
3. Random Forest.
```
Accuracy of  Random Forest :  83.85
```
4. Logistic Regression
```
Accuracy of Logistic Regression: 61.51
```

Performance Comparison:

- Among the four models, the Random Forest Classifier achieved the highest accuracy at 83.85%, suggesting it was the most effective at predicting the correct labels for the given dataset.
- The Decision Tree Classifier had a moderate accuracy of 67.46%, performing better than the Gaussian Naïve Bayes and Logistic Regression but not as well as the Random Forest.
- Gaussian Naïve Bayes and Logistic Regression had lower accuracies of 60.02% and 61.51%, respectively, indicating they were less effective on this dataset compared to the other models.

7. Conclusion:

● Most Popular Features: The most notable features influencing crop yield are the type of crop, area under cultivation, production quantity, annual rainfall, fertilizer usage, and pesticide usage.

• Positive Aspects: Random Forest proved to be highly effective for crop yield prediction, achieving the highest accuracy among all models tested. This indicates its strength in handling diverse agricultural data and capturing complex relationships between features.

• Negative Aspects: Decision Tree, Gaussian Naïve Bayes, and Logistic Regression demonstrated lower accuracy, suggesting limitations in their ability to predict crop yields accurately. These models may struggle with the variability and complexity of agricultural data.

• Model Performance:

Decision Tree Classifier: Achieved an accuracy of 67.46%, indicating moderate performance with a tendency to overfit the training data.

Gaussian Naïve Bayes Classifier: Achieved an accuracy of 60.02%, reflecting its assumptions about feature independence which may not hold true for agricultural data.

Random Forest: Achieved an accuracy of 83.85%, highlighting its robustness and ability to generalize well from the training data.

Logistic Regression: Achieved an accuracy of 61.51%, showing limited performance likely due to the linear nature of the model not capturing complex relationships.

• Best Performing Model: The Random Forest model was the best-performing model, with the highest accuracy of 83.85%. This makes it a reliable choice for crop yield prediction, capable of leveraging the full spectrum of available features.

Overall, this analysis underscores the importance of model selection in crop yield prediction. The Random Forest model, with its superior performance, can significantly aid in accurately estimating crop yields. This can help optimize agricultural planning, resource allocation, and ultimately improve food production efficiency.